Pattern Recognition Letters 63 (2015) 59–65

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Spatial co-training for semi-supervised image classification[☆]

Yi Hong [a,b], Weiping Zhu [a,*]

[a] International School of Software, Wuhan University, Wuhan 430079, PR China
[b] @WalmartLabs, San Bruno, CA, USA

ARTICLE INFO

Article history:
Received 23 December 2014
Available online 23 June 2015

Keywords:
Co-training
Semi-supervised learning
Image classification

ABSTRACT

Co-training is a famous learning algorithm used when there are only small amounts of labeled data and large amounts of unlabeled data, but it has a limited application in image classification due to the unavailability of two independent and sufficient representations of a single image. In this paper, we propose a novel co-training algorithm, in which these two independent and sufficient representations are automatically learned from the data. We call it as the spatial co-training algorithm (SCT). The main idea of the SCT algorithm is to divide an image into two subregions and consider each of them as an independent representation. In the SCT algorithm, the division of the image is firstly learned by an EM style algorithm on small amounts of labeled images, and finally relearned by a co-training style algorithm on many confident unlabeled images; while the classification of the image is performed jointly with the division of the image. We validate the SCT algorithm by experimental results on several image sets.

1. Introduction

Recent years have witnessed increasing interest in image classification [1–5]. This interest resulted in many effective approaches that progressed the computer vision field very fast. For example, the classification accuracy on the Caltech-101 dataset has climbed up from under 20% in [2] to almost 90% [6]; and the number of image categories that can be processed has also increased from around 100 in the Caltech-101 dataset [2] to almost 10 thousands in the ImageNet dataset [7].

Despite these progresses, current approaches are still sensitive to their amounts of labeled images and their high accuracies rely on large amounts of labeled images, sometimes more than 100 per category. A typical example is the experimental result on the Caltech-101 dataset: accuracies of kNN [4], SVM [8] and Random Forests [3] are around 75%, 78% and 88% respectively if 30 images per category are labeled; but their performance degrades a lot, just around 65%, 70% and 70%, if this number reduces to 15. It is in fact a big challenge for a variety of classifiers to achieve satisfactory results when only small amounts of images are labeled [9,10]. On the one hand, human labeling is time-consuming and boring, and how many images per category should be labeled still remains an unsolved problem. On the other hand, there is a huge number of unlabeled images that can be downloaded from the Internet. It is therefore quite promising to improve the accuracy of the classifiers by the way of making use unlabeled images.

Many effective approaches have been proposed so far to improve accuracies of classifiers by making use of large amounts of unlabeled data [11–13]. Among them, the co-training algorithm assumes that each example has two different representations; in addition, these two representations are conditionally independent and sufficient enough for a good classification [11]. It thus iterates the following two steps: (1) to learn a separate classifier with respect to each representation of all labeled data; (2) to augment existing labeled data by most confident predictions of each classifier on unlabeled data. The co-training algorithm has been widely applied in the fields of email classification [14], web page mining [11] and visual tracking [15]. However, we have not observed many of its applications in the field of image classification, apart from [16] where the content of the image and its tags are used as two independent representations for web image classification and [17] where contour and skeleton are considered as two complementary representations for shape retrieval. As pointed out in [18], this is because it is usually hard to obtain such two independent and sufficient representations of a single image. Nevertheless, several recent studies have shown that the above assumption is too strong and can be relaxed a lot [19–21]. In [20], the authors proved that a weaker expanding property on the underlying distribution of the data is enough for the success of co-training algorithm. They suggested that co-training algorithm should work well if there are at least some cases when the classifier on one representation makes confident decisions while the classifier on the

☆ This paper has been recommended for acceptance by D. Dembele.

* Corresponding author: Tel.: +86 27 68778720.
E-mail address: wpzhu@whu.edu.cn (W. Zhu).

http://dx.doi.org/10.1016/j.patrec.2015.06.017
0167-8655/

- **Input**
  - A set L of labeled training examples
  - A set U of unlabeled examples
- **Process**

  Create a pool U' of unlabeled examples by choosing u' examples at random from U

  Loop for k iterations

  (1) Train a classifier F_1 on L that considers only the x^1 portion of x

  (2) Train a classifier F_2 on L that considers only the x^2 portion of x

  (3) Label p positive and n negative on which F_1 has most confident predictions

  (4) Label p positive and n negative on which F_2 has most confident predictions

  (5) Add these self-labeled examples to L

  (6) Randomly choose $2p + 2n$ examples from U to replenish U'

Fig. 1. The co-training algorithm by [11].

other representation does not have much confidence in its own decision. The above weaker expanding property of co-training algorithm has been well demonstrated in [19] where the authors showed that a random split of a feature set usually makes co-training algorithm success and in [21] where the authors proposed an elegant algorithm to automatically decompose a single feature set into two complementary subsets as inputs of co-training algorithm.

Inspired by the above weaker expanding property, in this paper we propose a novel co-training algorithm, in which these two independent and sufficient representations are roughly learned from images. Another inspiration of the proposed algorithm comes from the property of image classification task itself. As for image classification, the information contained in a whole image is usually redundant. In this case, a subregion of the whole image is usually sufficient enough for the classifier to make a confident prediction on it. For example, if our task is to do human vs. non-human image classification, and if we have already known there is a face in the image, we can make sure that the image belongs to the human image, without further studying if there are two legs or not. We call the proposed algorithm as the spatial co-training algorithm (SCT).

2. The spatial co-training algorithm

We start the discussion from traditional co-training algorithm. Let $L = \{(x_i^1, x_i^2, y_i), i = 1, \ldots, \ell\}$ denote the set of labeled examples and $U = \{(x_i^1, x_i^2), i = \ell + 1, \ldots, \ell + u\}$ denote the set of unlabeled examples, where x_i^1 and x_i^2 are two different representations of the same example x_i, co-training algorithm thus performs to find the labels $\{y_i, i = \ell + 1, \ldots, \ell + u\}$ by making use of both representations cooperatively. Fig. 1 shows the framework. Co-training algorithm assumes that each example has two different representations and these two representations are sufficient enough for a good classification. Note it is sometimes quite natural to obtain such two representations. For example in web page classification problems, both words in web page and words used in another page that linked to the page are discriminative enough for classifying academic and non-academic personal web pages, therefore the histogram of all words appearing in web page and the histogram of all words appearing in other pages that linked to the page can be used as two different representations [11]. But it is really hard to obtain such two complementary representations for many other problems such as image classification. In this paper, we propose a novel co-training algorithm that learns such two representations automatically from images.

2.1. Problem definition

The proposed algorithm is based on bag-of-words model [22] that represents each image as a histogram of its local image patches. In particular, bag-of-words model performs to: (1) extract a collection of local descriptors from the images; (2) quantize them as indexes; (3) and represent each image as the histogram of indexes of its local image patches. Many visual codebook learning algorithms have been proposed so far, and in this paper we employ the k-means clustering algorithm because of its simplicity and wide applications.

Given a set of labeled images $\{I_k, k = 1, \ldots, \ell\}$, we extract local descriptors densely from each image and express them as a matrix F:

$$F = \{f_{ij} | i = 1, \ldots, w;\ j = 1, \ldots, v\}, \tag{1}$$

where f_{ij} is the local descriptor of local image patch (i, j) that is usually a vector, w and v are the height and width of the local description matrix. To simplify our notations, in the following paragraphs we assume that all images have the same size and thus have the same values of w and v of the matrix F, even though recent study in [23] can deal with images with different size conveniently. Then we apply the k-means clustering algorithm to quantize these local sift descriptors into indexes and rearrange indexes from one image as a matrix C:

$$C = \{c_{ij} | i = 1, \ldots, w;\ j = 1, \ldots, v\}, \tag{2}$$

where $c_{ij} \in \{1, \ldots, K\}$ is the visual code of the local image patch (i, j) and K is the codebook size. As illustrated in the Introduction, the basic idea of the proposed algorithm is to partition an image into two subregions and consider each of them as an independent representation. Now suppose that I_1^s and I_2^s are a partition of the image I that corresponds to a partition C_1^s and C_2^s of all visual codes in C:

$$C_1^s \bigcap C_2^s = \emptyset; \quad C_1^s \bigcup C_2^s = C, \tag{3}$$

we calculate the histograms of visual codes in both I_1^s and I_2^s as:

$$h_{1,k} = \frac{\sum_{(i,j)\in C_1^s} \delta(c_{ij}, k)}{\sum_{(i,j)\in C_1^s} \mathbf{1}}; \quad h_{2,k} = \frac{\sum_{(i,j)\in C_2^s} \delta(c_{ij}, k)}{\sum_{(i,j)\in C_2^s} \mathbf{1}}, \tag{4}$$

after normalized for $k = 1, \ldots, K$; and $\delta(a, b) = 1$ if $a = b$, and $\delta(a, b) = 0$ if $a \neq b$. According to (4), each division (C_1^s, C_2^s) of the image will lead to a representation pair (h_1, h_2) of the image:

$$\left(C_1^s, C_2^s\right) \Rightarrow (h_1, h_2), \tag{5}$$

where both h_1 and h_2 are the histograms with K bins. Consider the overall number of divisions of each image is around wv that is too large to be processed, in this paper we restrict the division of the image as a vertical line. In this case, the partition of C at the position d is $(C_{1:w,1:d}, C_{1:w,d+1:v})$, and we simplify it as $(C_{1:d}, C_{d+1:v})$ together with their histogram representation pair as $(h_{1:d}, h_{d+1:v})$. Therefore, the candidate pool of all possible representation pairs is:

$$H = \{(h_{1:d}, h_{d+1:v}) | d = 1, \ldots, v - 1\}. \tag{6}$$

Based on the above notations, the proposed algorithm aims to find good divisions $\{d_1, \ldots, d_\ell, d_{\ell+1}, \ldots, d_{\ell+u}\}$ of both labeled and unlabeled images such that the co-training algorithm using $\{(h_{i,1:d_i}, h_{i,d_i+1:v}) | i = 1, \ldots, \ell; \ell + 1, \ldots, \ell + u\}$ as inputs performs well on unlabeled images. The proposed algorithm has two separate steps: (1) first it learns divisions of labeled images $\{d_1, \ldots, d_\ell\}$ by an EM style algorithm; (2) then it learns divisions of unlabeled images $\{d_{\ell+1}, \ldots, d_{\ell+u}\}$ by a co-training style algorithm. In the following paragraphs, we will go into details of describing these two steps Fig. 2.